

# **ANALYSIS OF A TEACHER PAY-FOR-PERFORMANCE PROGRAM: DETERMINING THE TREATMENT EFFECT AND OVERALL IMPACT**

**By John T. White and Jill G. Leandro,<sup>1</sup> SAS EVAAS  
SAS Institute, Inc.**

## **SECTION 1: PURPOSE OF THIS STUDY**

Teacher pay-for-performance or incentive programs are becoming more common as a part of educational policy and practice across the nation. In many of these programs, policymakers determine the influence that teachers have on their students' academic progress, and with this knowledge, the teachers are rewarded according to the program's criteria.

With the rise in popularity of these programs, there have been corresponding efforts to determine their effectiveness. More specifically, there is an interest in whether or not the programs have an impact on student-level achievement gains. This leads to two important questions. The first question is this: Is there a treatment effect? In other words, if a teacher receives an award, does that have any impact on his or her future performance? This can be thought of as the treatment effect of the pay-for-performance program since teacher performance in this context is a positive influence on student achievement gains. This study uses a similar approach to others for determining the treatment effect on a performance pay program by using a regression discontinuity analysis.

The second important question for this study is: What is the overall impact of the pay-for-performance program? Has the district benefited *overall* since implementation of its pay-for-performance program? This can be assessed in several ways, but this study will analyze teacher mobility trends in the district as well as consider how the district performs relative to the rest of the districts in its state.

Thus, this study explores whether any treatment effect exists and, if so, how that impacts student-level progress. To do this, a three-year-old teacher pay-for-performance program from a large urban district was analyzed. Instead of comparing this school district with a comparable district that does not have a pay-for-performance program, the focus of the study will be to determine what impact the actual award has had on students taught by teachers in that district.

Section 2 describes the data used for this study. Section 3 describes the actual analysis. Results are reported in section 4, and the conclusions are in section 5.

---

<sup>1</sup> See page nine for the authors' biographical information.

## SECTION 2: VALUE-ADDED MODELING AND DATA USED IN ANALYSIS

Houston Independent School District (HISD) employs a teacher pay-for-performance program that has used estimates of teaching effectiveness from the SAS® EVAAS® teacher value-added model in the three most recent school years. More specifically, after each school year was completed, teachers were awarded based on their influence on student progress exhibited with respect to the district average. While these teachers were also awarded based on school-level value-added results (along with non-core teachers and other school personnel), the focus of this study is on the *individual* teacher component of the pay-for-performance program for grades three through eight.

For the analysis, there were three years of data for estimating a teacher's influence on the growth of students in their classrooms: 2006 – 2007, 2007 – 2008, and 2008 – 2009. (For brevity in what follows, school years are identified by their *ending* year; i.e., the three years are 2007, 2008, 2009.) The analysis employed scores on the state tests in Math and Reading from the 2005 to 2009 school years. The student-level data only included students who tested in English.

Before application to any modeling, student-level scores were converted to normal curve equivalents (NCEs). In contrast to percentile rankings, NCEs provide an equal-interval scale such that the difference between the 80<sup>th</sup> and 90<sup>th</sup> NCE is equivalent to that between the 40<sup>th</sup> and 50<sup>th</sup> NCE, which is an important distinction for measuring student academic progress. The distribution of *state* scale scores from 2006 was used to create a mapping between student scale scores and NCEs. Using this mapping, all of the district's student-level data were converted to state NCEs with a base year of 2006, which allows each student's score to be relative to the student achievement of the state in 2006. Once the student scores had been mapped, they were then used in a multivariate longitudinal model, which exploits the relationships among the scores in a student's entire informational array to dampen the effects of measurement error around any single test score.

### Part I: Is There a Treatment Effect?

These data provided estimated teacher effects on classroom gains that could address the first important question of this study. The teacher value-added reports provide a measurement of influence in terms of an estimated teacher gain in NCE units. The report also provides the standard error associated with the gain to determine the level of uncertainty around the estimate. The teacher gain is a *shrinkage estimate* such that the teacher is assumed to have the overall district gain observed in that subject/grade for that year until the weight of the evidence pulls the teacher's gain away from that average.

To ensure the comparability of the teacher's value-added estimates across school years, only terminal year teacher gains were used. Thus, when computing a teacher's 2007 value-added result, years 2005 – 2007 were used. Similarly, to obtain a teacher's 2008 value-added estimate, data from 2005 – 2008 were used; and for 2009 estimates, years 2005 – 2009 were used.

As the HISD pay-for-performance program awards teachers relative to the district average across grades in a given subject/year, not for a particular subject/grade/year, the modeling must also account for this distinction since the level of uncertainty can vary across different grades. For this reason, the HISD model utilizes an index, which is calculated in several steps. First, the district gain is subtracted from the teacher gain since these gains are centered on the district average. This is called the teacher effect. For teachers who teach a single grade for a given subject, the index is calculated by dividing the teacher effect by its standard error. For teachers who teach multiple grades in the same subject, their teacher effects are averaged across grades and the average teacher effect is divided by its appropriate standard error to find the index.

Dividing by the standard error, or level of uncertainty, allows these values to be ranked and compared. Using this process, each teacher could have an index from each of the years 2007 – 2009 in each subject. Of course, many will not have a complete set of indices because teachers leave the district or change to subjects other than math and reading. Regardless, all teachers with sufficient data were included in the analysis.

The HISD model awards the top two quartiles of these ranked indices, so long as they are above zero. (Note that while the indices for a particular subject/year have a mean of zero, the median, which is the award cutoff that defines the top two quartiles, may differ slightly from zero.) The award model has two levels of payment, based on whether the teacher is in the top quartile or second-highest quartile for each subject. Within each year, a teacher can be awarded for each subject that they taught. The analysis data consist of the teacher's index for each of the three possible years in math and/or reading as well as categorical variables indicating the quartile of the teacher's index for each of the three years in each of the subjects.

## **Part II: What is the Overall Impact of the Teacher Pay-for-Performance Program?**

To address the study's second important question, the study used the process described above as well as a separate analysis. Student scale scores were converted to state NCEs within year, rather than to the 2006 base year, to determine HISD's relative place in the distribution of scores with respect to the state. The change in achievement level from year-to-year is compared to determine if the district had an increase in student achievement since implementing this pay-for-performance program.

## **SECTION 3: MODELS AND ANALYSIS**

### **Part I: Is There a Treatment Effect?**

In order to address the study's first important question of treatment effect, a regression discontinuity model<sup>2</sup> was used to determine whether a treatment effect exists in HISD's pay-for-performance program through its use of a monetary award to qualifying teachers. This model utilizes the concept of a break point for subjects that were included into the treatment category. As is the case with this study, the break point can be a sharp cut in the data. In HISD, teachers are paid from the pay-for-performance program if they are in either the second-highest quartile or top quartile for the teacher index in a given subject/year. A narrow band of observations around the break point is equivalent to a randomized sample of observations. The index described in section 2 has units in terms of the number of standard errors away from the district average. A band of 0.25 standard errors around the break point forces all observations to be within 0.5 standard errors of one another. Because there is essentially little difference, if any, between these points except for the uncertainty around them, this selection can be thought of as equivalent to a randomized sample. As such, if it can be shown that the teachers who happened to be above this break point influence student gains more significantly in the next year than those teachers below the break point, then the pay-for-performance program has a positive impact on the student level gains that following year.

The regression discontinuity model fits two regression lines using the teacher value-added index for a particular year as the response and a previous year's index (one or two years previously) as the predictor variable. Only teachers in a narrow band ( $\pm h$  standard errors) around the break point are used. One line is fitted for teachers who are below the break point by no more than 'h' standard errors, and the other line is for teachers who are just above the break point by no more than 'h' standard errors. The two regression lines are then compared at the break point value to determine the treatment effect. If this difference proves statistically different from zero, then the null hypothesis of no measurable treatment effect is rejected.

### **Part II: What is the Overall Impact of the Teacher Pay-for-Performance Program?**

In order to address the study's second important question on overall impact, two different approaches were examined. First, because many teachers enter and exit the school district every year at HISD and the sample of teachers changed between 2007 and 2009, insight is provided into these teacher mobility trends. For instance, if teachers who stay in the pay-for-performance program longer tend to move into the award categories, then the program could be having a positive impact on the teacher's ability to influence student gains. Additionally, teachers who

---

<sup>2</sup> Van der Klaauw, W. (2008). Regression-discontinuity analysis: A survey of recent developments in economics. *Labour*, 22(2), 219-245.

leave the program should be examined to determine if there is a trend in their award categories as well.

As a second approach to the overall impact, another district-level value-added model was run to obtain the overall HISD mean NCE score for each grade and subject. This part of the study used student scores that were converted to state NCEs *within year*, rather than to a base year, so that it would be possible to evaluate HISD's performance relative to the state. For example, if the whole state increases its progress from one year to the next, then it would be of interest to evaluate how HISD's progress rate compares to the state progress rate. Since HISD's current ASPIRE Award program utilizing EVAAS data was incorporated into the district at the end of the 2006 – 2007 school year, the study used mean NCE scores averaged across grades for each subject from the 2006 – 2009 school years. The change from 2006 to 2007 was used as a baseline for the district, and HISD's progress each subsequent year was compared against that baseline to evaluate the district since its implementation of the current pay-for-performance program.

## **SECTION 4: RESULTS OF ANALYSES**

### **Part I: Is There a Treatment Effect?**

The results for the regression discontinuity analyses are given in Table 1. The response variable in each analysis is the teacher index for the later of the two years listed in the first column of the table. The predictor variable is the teacher index for the earlier year listed in column one. The break point is at the median of the predictor variable. The analyses use bands of  $\pm 0.25$ ,  $\pm 0.50$ , and  $\pm 0.75$  standard errors to test for a measurable treatment effect. The treatment effect, which is also the difference between the two regression lines at the break point for each band of data, is given in the treatment effect row of Table 1 along with its standard error below it. This information is used to obtain the p-value just below the standard error. Three levels of significance are noted in this table. P-values below 0.05 represent the tests with the most significant evidence to reject the null hypothesis of no treatment effect. The table also notes those treatment effects that are significant at the 0.10 level and the 0.15 level.

**Table 1: Regression Discontinuity Results for Math and Reading in All Year Transitions**

Years Examined		h=0.25		h=0.50		h=0.75	
		Math	Reading	Math	Reading	Math	Reading
2007 to 2008	Treatment Effect	2.49***	0.94**	1.12**	0.10	0.78*	0.14
	Standard Error	0.86	0.56	0.60	0.39	0.51	0.32
	P-value	0.005	0.097	0.065	0.807	0.123	0.653
	Observations	79	116	142	232	201	321
2007 to 2009	Treatment Effect	0.95	0.73	1.16*	0.69*	0.40	0.54*
	Standard Error	1.09	0.57	0.75	0.45	0.65	0.36
	P-value	0.386	0.205	0.127	0.123	0.537	0.132
	Observations	62	98	115	198	166	271
2008 to 2009	Treatment Effect	-1.08	0.40	-0.10	0.00	0.75*	-0.04
	Standard Error	0.86	0.48	0.57	0.33	0.50	0.28
	P-value	0.211	0.409	0.861	0.996	0.136	0.888
	Observations	86	127	160	213	241	307

Note: \* indicates statistical significance at the 15% level, \*\* at the 10% level, \*\*\* at the 5% level.

Significant differences are denoted in Table 1. Note that these significant differences are always associated with a positive treatment effect. In other words, if a teacher did receive an award by random chance in the first year, then he or she was more likely to have a *better* influence on students in the second year, indicating that the treatment/pay-for-performance could have been a likely factor in the improved performance.

Out of the 18 treatment effects, only 3 of them are negative, and they were not significant. 15 out of the 18 treatment effects were positive although not all were significantly positive. 8 out of the 15 positive were significantly positive at least at the 0.15 level. It appears that the largest positive treatment effects are in the 2007 to 2008 comparison. One conjecture is that this could have been caused by the district's emphasis on training regarding student growth measures and the use of value-added data during the 2007-2008 school year.

## **Part II: What is the Overall Impact of the Teacher Pay-for-Performance Program?**

Exploring the impact of the pay-for-performance program further, the teacher mobility trend analysis provided further evidence of a positive treatment effect. By definition, teachers with indices in the top two quartiles receive awards in any given subject/year, provided the index is positive. Thus, approximately 50% of the teachers in a given subject for that year receive a teacher reward based on their individual value-added assessment. This analysis sought a

response to questions on which types of teachers were likely to remain or leave the district: those who received an award or those who did not?

The study found that the teachers who leave the district tend to be the teachers who did not receive an award. While 50.36% of the teachers did not receive an award based on their performance in 2007, a much larger percentage of the teachers who were not in the data in either 2008 or 2009 did not receive an award in 2007: 59.17%. This trend is positive since these teachers are more likely to have less desirable impacts on student level achievement gains. A similar trend exists based on the 2008 awards. In 2008, 49.96% of the teachers did not receive an award based on their performance. Of the teachers who were not in the 2009 data, 57.18% did not receive awards in 2008. Again, this finding suggests a positive outcome from the pay-for-performance program.

Shifting focus to the teachers who remain in the district over time, the study suggests that these teachers have a slightly higher chance of obtaining an award than teachers who did not. For teachers who remained in HISD for all three years of the study, more than 50% received an award each year. More specifically, 54.86% of teachers who remained in the district for all three years obtained awards in 2007. For this same subset of teachers, this percentage was updated to 57.48% in 2008 and 54.11% in 2009. Since the expected percentage would be approximately 50% of the teachers, this finding suggests that teachers who remain in the district with access to this program are more likely to receive awards.

These differences are non-trivial, as the sample sizes were quite large. There were 2,202 teacher/subjects eligible for awards in 2007; 2,430 teacher/subjects in 2008; and 2,436 teacher/subjects in 2009. Of all of these teachers/subject observations, 1,070 of them were in each year from 2007 to 2009.

Using the district value-added model rather than the teacher model, Figure 1 illustrates how HISD performed relative to the rest of the state between 2006 - 2009. In this graph, each point represents the district's mean NCE in math and reading. As a reminder, this process uses intra-year state NCEs so that HISD can be evaluated against the state distribution of student scores. HISD implemented its current pay-for-performance program in 2007, so the change between 2006 and 2007 represents the baseline change from year-to-year that HISD experienced *prior* to implementing the current pay-for-performance program based on value-added data. The baseline change is compared to the change from 2007 to 2008 and 2008 to 2009, both of which represent a comparable one year change in HISD with the addition of the current teacher pay-for-performance program in place.

As illustrated in Figure 1, the rate of change has increased in both subjects. While the first year of implementation had the largest increase in rate of change for both subjects, the second year after implementation of the teacher pay-for-performance program still shows a larger rate of change for both subjects than the baseline year.

**Figure 1: District Mean NCE Across Years (Using State Intra-Year NCEs)**

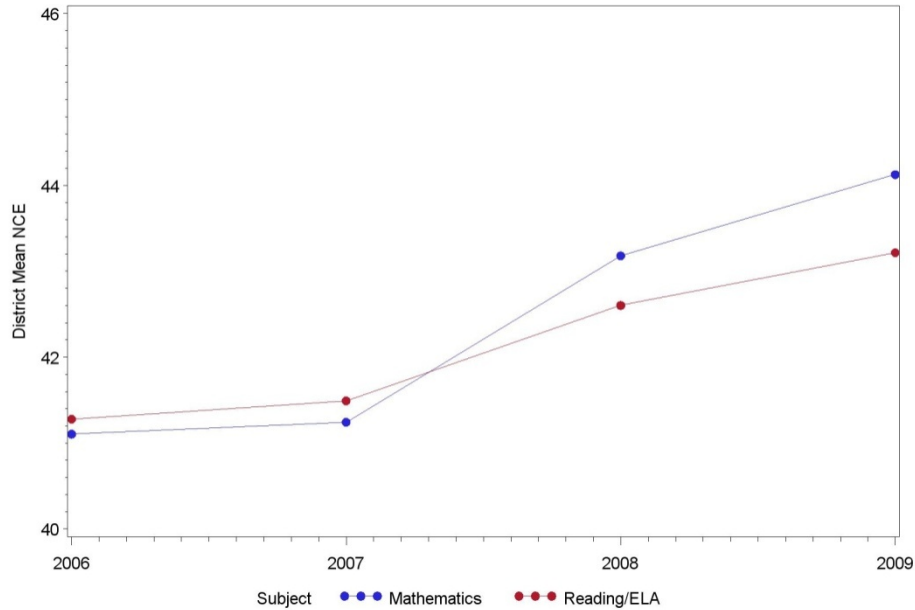


Table 1 shows the rate of change in each of the year transitions.

**Table 1: Rate of Change in Achievement from Year to Year**

Subject	Year Change	Rate of Change
Mathematics	2006-2007	0.139
Mathematics	2007-2008	1.938
Mathematics	2008-2009	0.946
Reading/ELA	2006-2007	0.212
Reading/ELA	2007-2008	1.112
Reading/ELA	2008-2009	0.613

Using the appropriate standard errors from these means, the standard errors of the year-to-year differences in rate of change are all smaller than 0.13. This indicates that all of the differences are statistically significant. Both 2007-2008 and 2008-2009 are significant increases over 2006-2007. In other words, after the implementation of the current pay-for-performance program and the availability of value-added analysis, HISD was able to show a significant increase from where it originally was prior to implementing this program. The magnitude of the positive cumulative effect of improvement relative to the state's distribution is quite impressive, especially for a district of the size of HISD.

## SECTION 5: CONCLUSIONS

Since implementing its pay-for-performance program, HISD has had a positive impact on both student-level achievement gains as well as overall achievement. The regression discontinuity



analysis showed that when a teacher receives an award, there is a positive impact on that teacher's *future* student-level gains. In other words, if a teacher receives an award in one year based on individual value-added assessments, then he/she is likely to perform better in terms of value-added influence on students in the following year. Although not all of the comparisons were significant, most all of the differences were positive. All of the significant differences had a positive effect on student-level gains. This treatment effect of the pay-for-performance program seemed to be more prevalent in math than reading. This could be due to the fact, long-known, that there is more measurable differentiation among math teachers than reading teachers.

Since HISD has implemented the pay-for-performance program, teachers who stayed in the district for longer periods of time are more likely to receive individual awards. Also teachers who left the district were more likely to be teachers who did not receive awards. In other words, the teachers who are less likely to have a positive influence on student-level gains leave, and the teachers who are more likely to have a positive influence on student level gains remain in the district.

Furthermore, the analysis shows that the rate of change in achievement level of students from year-to-year in HISD has significantly increased since the implementation of the pay-for-performance and value-added analyses.

These findings lead the authors of this study to conclude that the impact of the pay-for-performance program has been positive on the educational experience for students in HISD. Given the evidence in this study, such an impact will likely persist into the future by continuing to use the pay-for-performance program based on teacher-level value-added data.

### **Biographical information for the preparers of this document:**

John White, MS, has been an analytical consultant with the SAS EVAAS group for three years and is presently pursuing a Ph.D. in statistics at North Carolina State University. He has a Masters in Statistics and was selected as the recipient of the Gertrude M. Cox Academic Achievement Award for the Outstanding MS Candidate in 2007.

Jill Gentry Leandro, MPP, represents the SAS EVAAS group in educational policy. She has a Master in Public Policy degree from Harvard University's John F. Kennedy School of Government.

The SAS EVAAS group is a part of SAS Institute Inc. ®, Cary, NC.